Bavarian State Office for
Data Protection
Supervision

Data protection compliant Artificial intelligence

Status of the checklist: 24.01.2024
Version: Consultation status v0.9

Aim and content of this paper

The checklist contained in this document sets out requirements for the development and use of applications in the "artificial intelligence" category. Due to the constantly progressing development, adjustments to this checklist may become necessary - in particular for harmonization with German and European data protection positions on AI. The checkpoints listed should therefore not be regarded as exhaustive, but represent a good practice approach that can be used in the sense of a target/actual review. This document is dedicated to the question of which data protection requirements can be of central importance when using artificial intelligence.

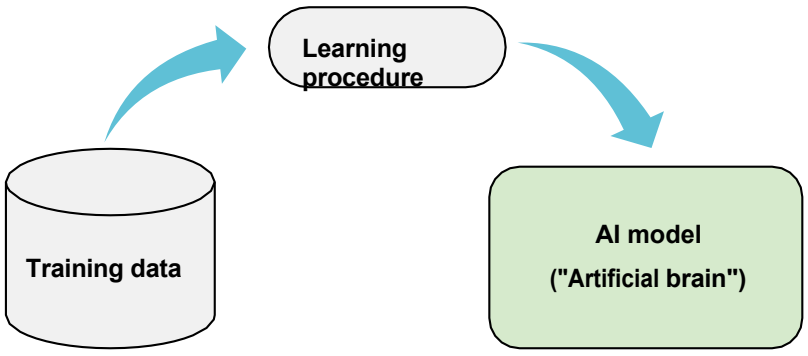
CONTENTS

A. Classification

Artificial intelligence applications essentially consist of an AI model (e.g. a large language model such as GPT). Such models are either trained from data - so-called machine learning - or explicitly modeled (e.g. rule-based systems, which are not so "hype" at the moment). Such models can also be thought of as an "artificial brain". In this respect, this checklist also takes a closer look at the training of artificial intelligence.

Artificial intelligence applications are operated in operational use, which means that the hosting scenario must be considered (e.g. on which cloud system or GPU computer an AI application is running). This sometimes raises similar questions to the use of cloud systems. When an AI application is used, input from an application environment (e.g. text in ChatGPT) is entered into an AI model, which sometimes has to be pre-processed for processing. The results of the AI model used (e.g. text output for ChatGPT or warning messages for pedestrian detection in a vehicle) are then also processed further. In this respect, this checklist also takes a closer look at the operation/use of artificial intelligence as an application.

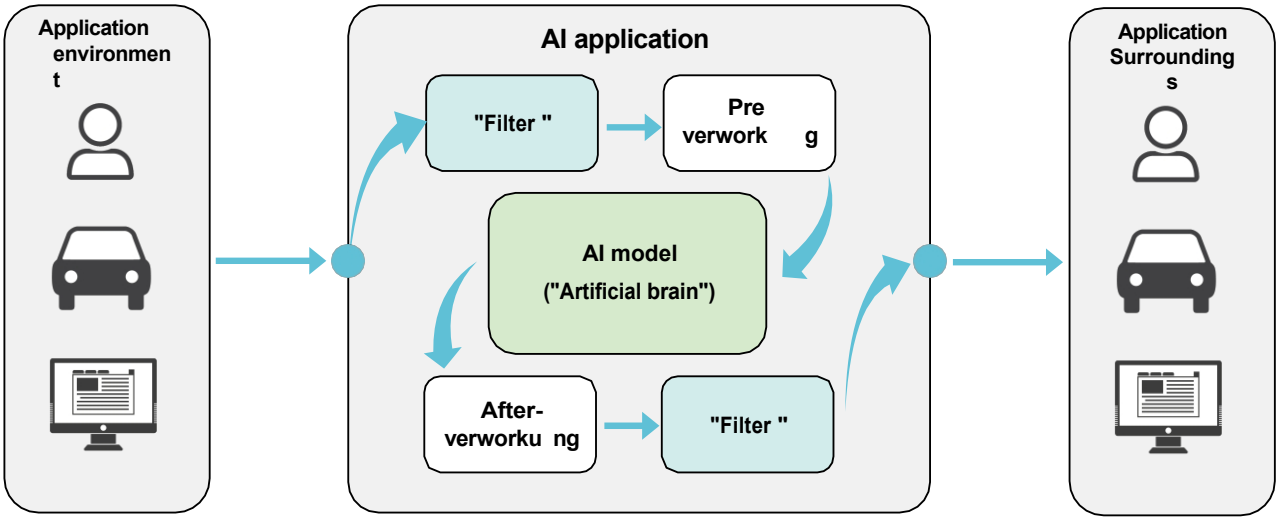
The following diagram illustrates the training of artificial intelligence:



Training of an AI model

An AI model always requires (at least for machine learning, but also for the explicit creation of rule systems) training data of a large scope, but in particular of very good quality (with regard to the application scenario). A learning method selected to match the AI model (e.g. gradient descent method to minimize a defined error class) attempts to train the essence of the training data (called "generalization") into an AI model, which should also be able to adequately process inputs not previously contained in the training data (e.g. new car model in a driver assistance system, individual conversation in ChatGPT).

The following diagram illustrates the operation and use of artificial intelligence:



Operation/use of artificial intelligence

A trained AI model is brought into the application ("application environment") by loading it into an IT environment (e.g. car for pedestrian recognition or GPU cluster ("AI-as-a-Service") for large language models such as ChatGPT) ("AI application"). Depending on the AI model, input/output data of the AI model is pre- and post-processed (e.g. transformation of audio signals into frequency ranges, coding/decoding between texts and number sequences in ChatGPT or image reduction in driver assistance systems). In addition, "filter systems" are sometimes used which, in the simplest case, simply filter out unwanted input such as the input of a hate message in a language model before it is entered. The results of AI models are also usually processed using appropriate filters, whether this is because the probability of recognizing a traffic sign is considered too low or because the results of a language model are also checked for unwanted output.

B. Training of AI models

The following checkpoints should be reviewed when creating AI models (possibly also as part of the data protection officer's monitoring activities):

- ☐ Definition and documentation of the type of AI technology to be implemented using the AI model (e.g. transformer architecture when creating a large language model including definition of the internal AI architecture such as layers and coupling as well as the number and initialization of the parameters)
- ☐ Evaluation of which training data is personal and which is not
- ☐ Inclusion of the training of AI models in the record of processing activities in accordance with Art. 30 GDPR. When training several AI models with different AI technologies, categories of personal data and, if applicable, recipients in a third country, it is advisable to include a separate entry for each.
- ☐ Check and document whether a data protection impact assessment (DPIA) must be carried out in accordance with Art. 35 GDPR. If necessary, process some of the checkpoints named in this section as part of the DPIA.
- ☐ Check whether AI training can also be carried out with anonymous data.
- ☐ Check whether AI training can also be carried out with synthetic or pseudonymous data.
- ☐ For synthetic training data: Check whether the algorithm for generating synthetic training data really creates anonymous outputs.
- ☐ Is there a legal basis for the use of personal training data?
- ☐ Consent is required for special personal data (e.g. health data) (or the exceptions in Art. 9 (2) GDPR apply).
- ☐ If the purpose of training an AI model is research, then check the relevant research privileges (the training of a large language model by a commercial company for product purposes is unlikely to fall under this category today).
- ☐ Documentation of all training data including sources (e.g. book databases or websites) in the sense of accountability according to Art. 5 para. 2 GDPR.

- ☐ The training data was checked for data quality with regard to statistical distortions or biases and adjusted/adjusted if necessary.

- ☐ For volatile training data (e.g. websites of news portals that can change quickly): Complete and audit-proof storage of the web page content used for the training.

- ☐ Sorting out training data that would bring unauthorized content into a training process (e.g. websites with fake news, hate content, conspiracy theories, ...). To this end, creation of a concept (in accordance with Art. 5 Para. 2 GDPR), according to which criteria the data is separated out.

- ☐ Cleansing of personal data not required for the training from the training data (e.g. credit card numbers, (e-mail) addresses, names, ...), taking into account country-specific coding (e.g. addresses are spelled differently in many countries).

- ☐ In addition to training/test data, validation data is also available, which is used to check the quality of the AI model and is not part of the training process.

- ☐ Creation of a risk model (next section) with which the quality of a learning procedure within the meaning of Art. 5 (2) GDPR can be demonstrated.

- ☐ Assessment and proof of whether or not an AI model created has a personal reference (has consequences for the transfer of an AI model and, if necessary, the need for a legal basis to be able to work on the personal AI models).

- ☐ Calculation and documentation of metrics from the risk model that can be used to demonstrate adequate containment of data protection risks (Art. 5 (2) GDPR). The search for suitable metrics is currently (also) still the current state of research.

- ☐ Implementation of the information obligations according to Art. 12 ff GDPR.

- ☐ Ensuring that requests for information in accordance with Art. 15 GDPR are also taken into account for requests for training in AI models in data protection management.

- ☐ In the case of a specific request for information in accordance with Art. 15 GDPR in relation to a personal AI model, it is checked - depending on the AI technology - whether personal data can be determined directly in the AI model or whether it can possibly only be derived from an AI model with additional information (e.g. specific prompt in the case of a large language model). In case of doubt, this additional information must then be requested from the data subject.

- [ ] Ensure that data subjects' rights to rectification under Art. 16 GDPR, to erasure under Art. 17 GDPR, to restriction of processing under Art. 18 GDPR, to data portability under Art. 20 GDPR and to object under Art. 21 GDPR with regard to AI are also taken into account in data protection management. Feedback deadlines for applicants must be observed.

- [ ] In the case of a deletion request in accordance with Art. 17 GDPR in relation to a person-related AI model, it is checked - depending on the AI technology - whether personal data can be determined directly in the AI model or whether it can possibly only be derived from an AI model with additional information (e.g. specific prompt in the case of a large language model). If deletion in an AI model is technically possible without affecting the overall model, the deletion process must also be carried out. If, on the other hand, personal data can only be determined from an AI model using additional information (e.g. prompts), one option for technical deletion is to use post-training to implement the specific personal AI output to be deleted by adjusting the internal (probability) parameters.

- [ ] When commissioning a service provider to (partially) take over AI training, check suitable guarantees and legal bases (e.g. contract for order processing, guarantees for third-country transfers such as adequacy decision, standard contractual clauses with transfer impact assessment or EU-US Data Privacy Framework). In particular, ensure that the data recipient does not use possible training data for its own purposes or at least that suitable legal bases and information obligations are complied with in the event of a change of purpose.

- [ ] If there is an obligation to carry out a DPIA in accordance with Art. 35 GDPR: Residual risk assessment based on the risk model and, if necessary, consultation of the competent data protection supervisory authority in accordance with Art. 36 GDPR if the risks to the rights and freedoms of the persons affected by the training remain high

- [ ] If the AI model is adapted during operation (e.g. by integrating some up-to-date websites), the respective model statuses including the respective training data must be stored in an audit-proof manner and taken into account in risk modeling in particular.

C. Assessment of risks with AI

Both the generation of AI models and their operation in an application environment (e.g. evaluation of customer letters in an insurance company, decisions as to whether a car should initiate emergency braking or not, generation of a political speech using ChatGPT, ...) are associated with risks. When processing personal data, the GDPR addresses the risks to rights and freedoms in such a way that these must be determined using an objective method (EW76) and mitigated with effective measures (Art. 25 GDPR, Art. 35 GDPR where applicable). For this reason, every controller and every processor whose processing of personal data is used either to generate AI models or to use them in AI applications must be aware of the specific risks associated with this and demonstrate these by means of suitable documentation in accordance with the accountability obligation under Art. 5 para. 2 GDPR. In the obligation to carry out a DPIA in accordance with Art. 35 GDPR, the specific high risks of artificial intelligence represent the core of the impact assessment.

For some years now, both research and increasingly regulatory/normalization bodies have been addressing the question of how artificial intelligence can be used for the benefit of data subjects and how the associated risks can be contained or at least reduced. This subject area is also referred to as "trustworthy AI" and can be used as a starting point for the formulation of data protection risks under the GDPR.

Creation of a risk model

- ☐ Determination and documentation of which of the following protection goals of an AI application are relevant for the specific scenario. Data protection risks then result from the deviation of a complete achievement of the respective protection goals ("data protection risk model"). Detailed justification if a protection goal is not considered relevant. These can be, for example, *based on the ethics guidelines for trustworthy AI*[1] of the European Commission:

  - ☐ "**Fairness**" in the sense that there are no unjustifiable risks of discrimination or unequal treatment.

  - ☐ "**Autonomy and control**" in the sense that there are opportunities to intervene in the operation of an AI application or that decisions with legal effect are not made without human control.

  - ☐ "**Transparency**" in the sense that, on the one hand, the data subjects are informed about the use of their personal data when training AI models and, on the other hand, that AI models and AI applications must be verifiable in terms of accountability. Also

[1] Available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

also that AI applications must be recognized as such for those affected in the interaction (e.g. chatbots or adaptation of audio signals to imitate a speaker).

- ☐ "**Reliability**" in the sense that an AI model or an AI application must, on the one hand, fulfill its intended purpose.
  The purpose of the system is fulfilled within tolerable error limits and that it is protected against visible manipulation (so-called adversial attacks, e.g. by means of prompt injection in large language models or interference in the recognition of traffic signs by special "stickers"). Hallucination (or confabulation) in large language models, in which incorrect output can sometimes appear undetected in an otherwise fluently formulated output, also belongs in this area.

- ☐ "**Safety**" in the sense that unwanted technical faults ("safety" such as hardware errors due to insufficient working memory), but above all unauthorized access/modifications ("security" such as the manipulation of training data during AI model generation or manipulation of "filters" that can be misused as a censorship mechanism) are effectively prevented.
  can be used.

- ☐ "**Data protection**" in the sense that, in addition to a legal basis for the creation of AI models and the operation/use of AI applications, the rights of data subjects and other
  compliance requirements of the GDPR must be implemented. This also includes the change of purpose of input data to an AI application by an AI operator for its own purposes. Important: Data protection risks under the GDPR can also include the above-mentioned risks of deviation from "fairness", "autonomy and control", "transparency", "reliability" and "security", insofar as personal data plays a role in these protection objectives.

- ☐ The data protection risk model must be documented and regularly checked to ensure that it is up to date and complete.

D. Use of an AI application

In an AI application, a trained AI model (which could also have been created by a third party) is used. To do this, it is loaded into an application environment on hardware that is usually specialized. The AI model is used by means of input values provided by a human, a mobile object (e.g. a car) or other software via an interface (e.g. a bank's chat system) ("application environment"). Outputs from the AI model are then returned to an application environment for further processing. Due to the very high hardware requirements of large language models such as ChatGPT, an AI application is often operated by a (cloud) service provider, with which it interacts either via a web interface or software interface.

Note: Securing the rights of data subjects in particular still poses a challenge for some AI applications (e.g. large language models). The question of data protection responsibility is important for controllers who use AI applications from large cloud providers as AI users. Before using an AI, it should be clarified whether the safeguarding of data subject rights relating to the AI model in an AI-as-a-service scenario is the responsibility of the AI provider, who may create an AI model themselves and only offer its use as a service and must then take care of the data subject rights themselves or is located with the client.

- ☐ Definition and documentation of the type of application to be implemented using AI technology (e.g. use of a large language model to create a chatbot for a bank or to assess the criticality of a customer complaint at a mail order company).

- ☐ Determining whether an in-house AI model should be operated in an in-house AI application (possibly on a service provider's hardware) or whether an AI application should be used with an AI provider (via a web interface or interface) that has complete control over the AI model, pre- and post-processing and the filter systems.

- ☐ Determine whether the AI model (your own or that of the AI provider) is in itself personal data. If so, the legal basis for the processing on the AI model must be examined (as a rule, a balancing of interests pursuant to Art. 6 para. 1 lit. f GDPR will be applicable).

- ☐ Inclusion of the use of an AI application in the record of processing activities in accordance with Art. 30 GDPR. If an AI system is used for several purposes (and possibly also for different categories of data subjects), it is advisable to make a separate entry for each purpose.

- ☐ Check and document whether a data protection impact assessment (DPIA) must be carried out in accordance with Art. 35 GDPR. If necessary, process some of the checkpoints named in this section as part of the DPIA.

- ☐ Determining and documenting which categories of personal data should be entered into an AI application as input data. In addition, definition and documentation of a legal basis.

- ☐ When entering special personal data in accordance with Art. 9 GDPR in an AI application, consent must be obtained (or the exceptions of Art. 9 para. 2 GDPR must be checked).

- ☐ Creation of a risk model (previous section) with which the data protection risks in the specific deployment scenario of the AI application can be mapped and verified.

- ☐ Calculation and documentation of metrics from the risk model that can be used to demonstrate adequate mitigation of data protection risks when using standard input data and the resulting output data (Art. 5 (2) GDPR). The search for suitable key figures is currently (also) still the current state of research, but at least a few test runs should be carried out, evaluated and documented.

- ☐ Definition and documentation of how to deal with risks from the risk model that arise in particular from faulty reliability (e.g. so-called hallucinations in large language models or the one-in-a-million events in autonomous vehicles in road use) and which sometimes cannot be determined by using suitable metrics (by means of test runs of inputs and outputs), but which can nevertheless occur sporadically in live operation.

- ☐ Implementation of the information obligations pursuant to Art. 12 ff GDPR (even if an AI-as-a-Service service is used).

- ☐ Ensuring that requests for information in accordance with Art. 15 GDPR also apply to requests regarding the use of
  AI applications must be taken into account in data protection management. In particular, the use of specific AI usage scenarios for personal data concerning the requesting person must be taken into account.

- ☐ In the case of a specific request for information in accordance with Art. 15 GDPR in relation to a personal AI model, it is checked - depending on the AI technology - whether personal data can be determined directly in the AI model or whether it can possibly only be derived from an AI model with additional information (e.g. specific prompt in the case of a large language model). In case of doubt, this additional information must then be requested from the data subject.

- ☐ Ensure that data subjects' rights to rectification under Art. 16 GDPR, to erasure under Art. 17 GDPR, to restriction of processing under Art. 18 GDPR, to data portability under Art. 20 GDPR and to object under Art. 21 GDPR with regard to AI are also taken into account in data protection management. Feedback deadlines for applicants must be observed.

- ☐ In the case of a deletion request in accordance with Art. 17 GDPR in relation to a person-related AI model, it is checked - depending on the AI technology - whether personal data can be determined directly in the AI model or whether it can possibly only be derived from an AI model with additional information (e.g. specific prompt in the case of a large language model). If deletion in an AI model is technically possible without affecting the overall model, the deletion process must also be carried out. If, on the other hand, personal data can only be determined from an AI model using additional information (e.g. prompts), one option for technical deletion is to use post-training to implement the specific personal AI output to be deleted by adjusting the internal (probability) parameters.

- ☐ Ensure that the data recipient of an AI-as-a-Service scenario does not use possible input data or output data from AI models and AI applications for its own purposes (e.g. retraining, filter improvement, marketing, etc.) or at least that suitable legal bases and information obligations are complied with in the event of a change of purpose. If necessary, AI applications must be specially commissioned or configured for this purpose.

- ☐ If there is an obligation to carry out a DPIA in accordance with Art. 35 GDPR: The company data protection officer must be involved. (Residual) risk assessment based on the risk model and, if necessary, consultation with the competent data protection supervisory authority in accordance with Art. 36 GDPR if there are still high risks to the rights and freedoms of the persons affected by the training.

- ☐ An approval test is carried out before an AI application is used. This must be documented.

- ☐ The use of AI applications is included in the data protection training program.

- ☐ The use of AI applications must be logged as proof of appropriate risk containment. Depending on the risk model, both input and output data must be stored on a secure log server with time stamps, taking strict account of the purpose limitation.

- ☐ When logging the use of AI applications, personal data that allows conclusions to be drawn about a specific employee is only stored in pseudonymous form using secure identification procedures.

- ☐ If the AI model is adapted during the ongoing operation of an AI application (e.g. by integrating some websites that are updated daily), this must be given special consideration in a risk assessment and release tests.

- ☐ The use of AI applications, including the data protection risk model, must be documented and regularly checked for up-to-dateness and completeness, taking into account the list of processing activities in accordance with Art. 30 GDPR.

Status of the checklist: 24.01.2024
Version: Consultation status v0.9

Current version for download:
https://www.lda.bayern.de/checkliste_ki

Publisher and contact:
Bavarian State Office for Data Protection Supervision (BayLDA)
Promenade 18 | 91522 Ansbach
www.lda.bayern.de | Tel.: 0981 180093-100
poststelle@lda.bayern.de